Track 1
Genome Skimming

OUTLINE

- Software needed
- Quality control
- Assemblers
  - Referenced-based (YASRA)
  - De novo (Velvet, SPAdes)

Computer Set-up

What do you need to have in your Computer?

- Internet connection
- Terminal/PuTTy access
- Azure cluster access
- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- Text editor (TextWrangler/TextEdit)
- Cyberduck (https://cyberduck.io/?l=en)
- Geneious
- Sequencher
- DOGMA account

What do you need to have in your cluster folder? SCRIPTS

- assemble_plastome.sh
- get_blasthit_seqs.pl
- new_fastq_cleaner.pl
- Spades folder (3 scripts)
- add_idline.pl
- Jellyfish folder (4 scripts)

What do you need to have in your cluster folder? DATASETS

- Raw reads dataset (Myspecies.fastq)
- Cleaned reads dataset – good (Myspecies_cutadaptcleaned.fq)
- Cleaned reads dataset - good
- Reference genome (Myreference.fsa)
- Reference genome Annotated (MyReferenceAnnotated.gb)
- Final choloroplast genome (Myspecies_CPgenome.fsa)
- Jellyfish output (Myspecies_CPgenome.fsa.coverage_20kmer.txt, Myspecies_CPgenome.fsa.coverage_20kmer.txt.problemareas.txt)
- YASRA output (Myspecies_Final_Assembly)
- SPAdes output (Myspecies_spades.cphit_seqs.fsa)
- Velvet output (Myspecies.cphit_seqs.fsa)

Quality Control

Quality controls steps (1)

- Study your dataset (Myspecies.fastq)
  - Is your data single or paired-end?
  - How many reads do you have?
  - What is the index that you used in your library preparation?
  - Are your reads of good quality? (FastQC)
  - What is the range of quality scores per base?
  - What is the most common quality score per read
  - Is your dataset GC biased?
  - How long are your sequences?
  - How many bad/empty sequences do you have in the dataset?

Quality controls steps (2)

- Remove adaptors (cutadapt commands within assemble_plastome.sh)

https://cutadapt.readthedocs.org/en/latest/guide.html

  - What is this script doing?
  - What is the sequence of the adaptor(s)?

- Remove low quality bases (new_fastq_cleaner.pl)
  - What is this script doing?
  - What are its parameters?

Quality controls steps (3)

- Study one of your cleaned datasets (Myspecies2.cutadapt.cleaned.fq)
  - How many reads are in your cleaned dataset?
  - Search ("**grep**" unix command) for this sequence in your cleaned reads: **GATCGGAAGAGCACACGTCTGAA**
    What is this sequence?

Quality controls steps (3)

Search ("**grep**" unix command) for
the other adaptor sequence
in your **cleaned reads**

Can you find it?
Why?

Quality controls steps (3)

- Study your “cleanest” dataset (Myspecies.cutadapt.cleaned.fq)
  - Are your reads of good quality? (FastQC)
  - What is the range of quality scores per base?
  - What is the most common quality score per read
  - How long are your sequences?
  - How many bad/empty sequences do you have in the dataset?

Assemblies

Reference-based assembly

YASRA

(http://www.bx.psu.edu/miller_lab/)

(https://umbc.rnet.missouri.edu/resources/How2RunYASRA.html)

Performs comparative assembly of short reads using a reference genome, which can differ substantially from the genome being sequenced.

Mapping reads to reference genomes makes use of LASTZ (Harris et al), a pairwise sequence aligner compatible with BLASTZ

De novo assemblers

- **Velvet**
- **SPAdes**

Put together short sequences by first, dividing them all into fragments of a given size (e.g. k-mer), and then utilizing de Bruijn graphs to establish possible paths that can connect all these k-mers together

(i.e. into contigs)

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph